Window-limited CUSUM for Online Change Detection and Performance Guarantees

Liyan Xie

Assistant Professor, The Chinese University of Hong Kong, Shenzhen (2021 – now)
Visiting Fellow, ECE, Princeton (2023 Fall)

Department of Mathematics and Statistics, Hunter College, CUNY
October 25, 2023

Joint work with George Moustakides and Yao Xie

Agenda

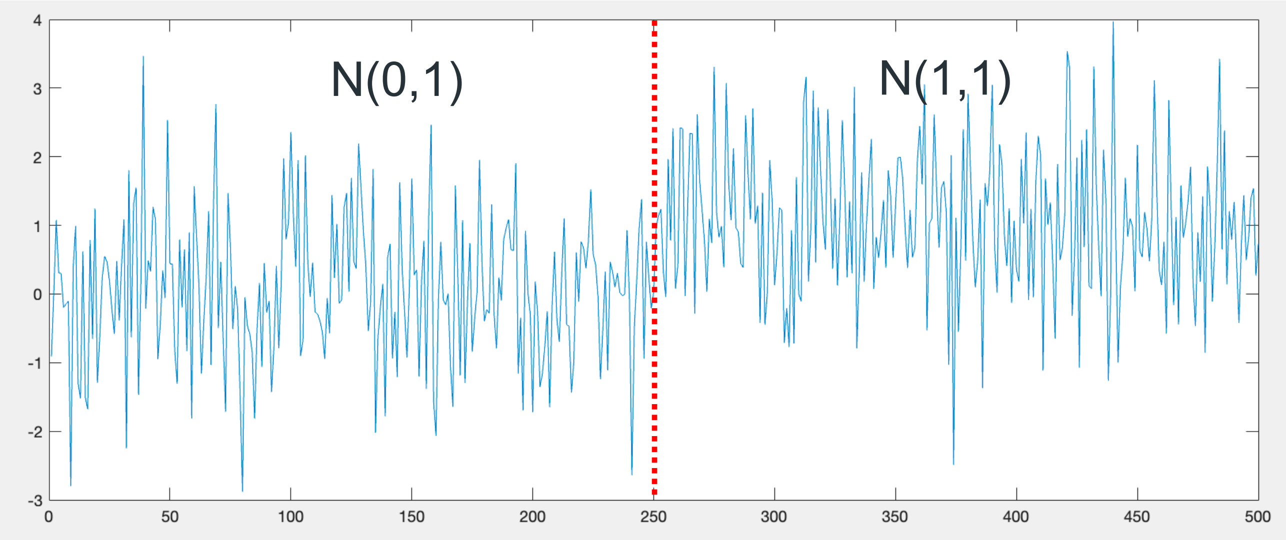
- Background on Sequential Change-point Detection
- A Newly Proposed Window-limited CUSUM Method
- Examples and Applications

Sequential Changepoint Detection

Sequential/online observations:

$$x_1, x_2, \ldots, x_\tau \overset{iid}{\sim} f_0 \qquad x_{\tau+1}, x_{\tau+2}, \ldots \overset{iid}{\sim} f_1$$

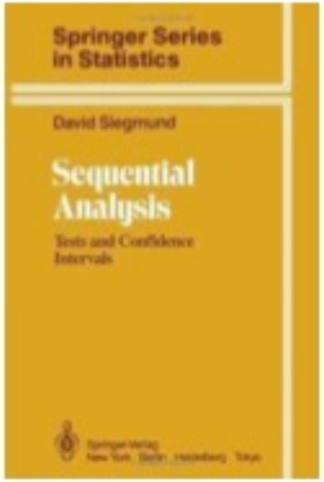
Change-point (unknown)

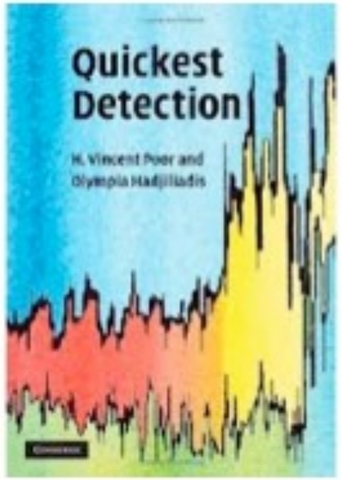


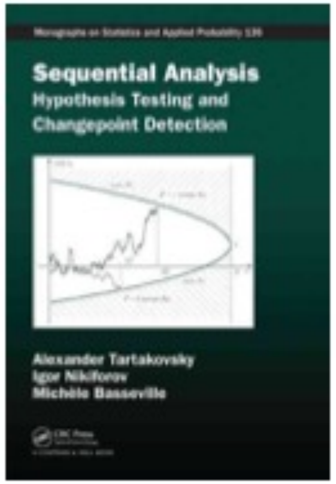
Mean shift

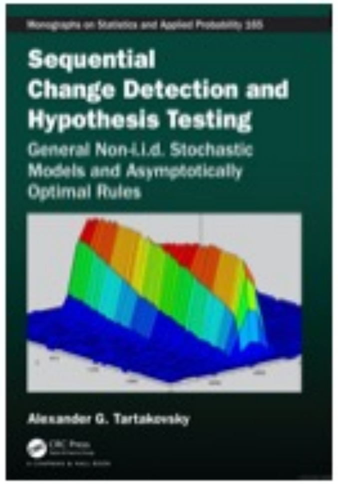
Multi-sensor and non-i.i.d

Single sensor and i.i.d









Page	Shiryayev, Roberts	Lorden, Moustakides	Siegmund	Basseville	Poor, Hadjiliadis	Tartakovsky, Nikiforovm Basseville	Tartakovsky
1954	1963 1966	1971 1986	1985	1994	2008	2014	2020

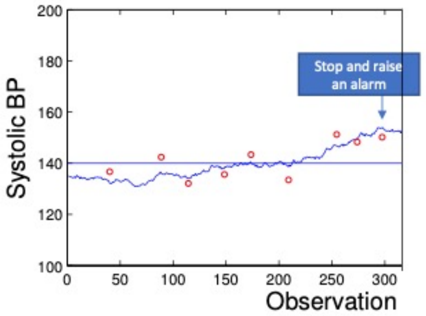
Quality control
Bomb production

Multi-arm bandit
Gambling

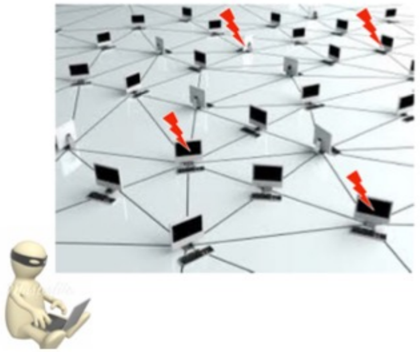
Sensor networks,
Social networks
Power networks,
Communication networks
Genetics

amount of data

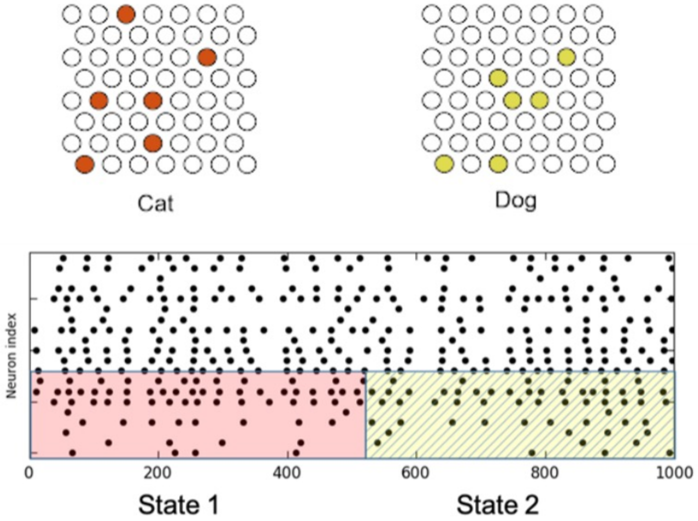
Application Examples



Bloop pressure monitoring

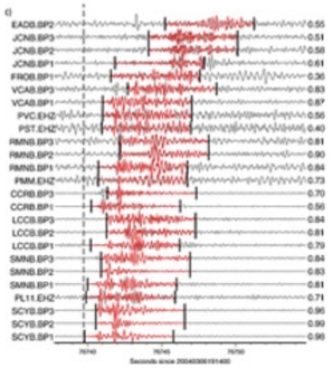


Anomaly detection

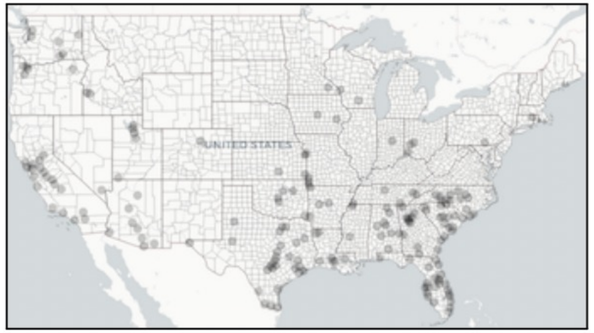


population code changes in neuronal networks

dense geophysical sensor array



Seismic sensor network



Covid hotspot detection

Sequential Changepoint Detection

The well-known Cumulative Sum (CUSUM) procedure:

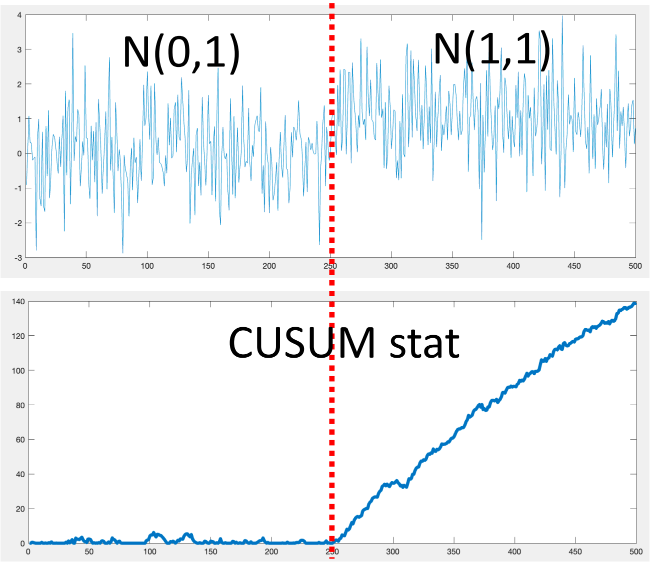
- CUSUM statistics: $S_t = \max_{k \leq t} \sum_{i=k+1}^{t} \log \frac{f_1(x_i)}{f_0(x_i)}$

  search over all past times

- A **recursive** implementation: ***Fully online and memory efficient***

$$S_t = \max\left\{S_{t-1} + \log \frac{f_1(x_t)}{f_0(x_t)}, 0\right\}, S_0 = 0$$

- **Stopping time:** $T_C = \inf\{t: S_t \geq \nu\}$



Limitation: need to fully specify f_0 and f_1 (almost impossible in practice)

Page, E. S. (1954). Continuous inspection schemes. Biometrika, 41(1/2), 100-115.

Generalized Likelihood Ratio test

- Recall CUSUM statistic:

$$S_t = \max_{k \le t} \sum_{i=k+1}^{t} \log \frac{f_1(x_i)}{f_0(x_i)}$$

- GLR statistic:

$$S_t^{GLR} = \max_{0 \le k < t} \sup_{\theta \in \Theta} \sum_{i=k+1}^{t} \log \frac{f_1(x_i; \theta)}{f_0(x_i)}$$

Stopping time:

$$T_G = \inf\{t : S_t^{GLR} \ge \nu\}$$

Maximum likelihood estimate

(Assuming the unknown post-change distribution $f_{1,\theta}$ is **parametrized** by θ)

(There are also non-parametric versions based on density ratio estimation.)

Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information theory*, *44*(7), 2917-2929.

Limitations of GLR test

- GLR statistic:

$$S_t^{GLR} = \max_{0 \le k < t} \sup_{\theta \in \Theta} \sum_{i=k+1}^{t} \log \frac{f_1(x_i;\theta)}{f_0(x_i)}$$

High computational complexity

- A window-limited version (computationally more efficient)

$$S_t^{wGLR} = \max_{t-w \le k < t} \sup_{\theta \in \Theta} \sum_{i=k+1}^{t} \log \frac{f_1(x_i;\theta)}{f_0(x_i)}$$

w must satisfy $\liminf w/\log\gamma > I_0^{-1}$ and $\log w = o(\log\gamma)$

Still high computational complexity and a large window is needed for strong false alarm requirements

Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information theory*, *44*(7), 2917-2929.

Optimality of CUSUM and GLR

- Goal: **min Detection Delay**
  s.t ARL $\geq \gamma$

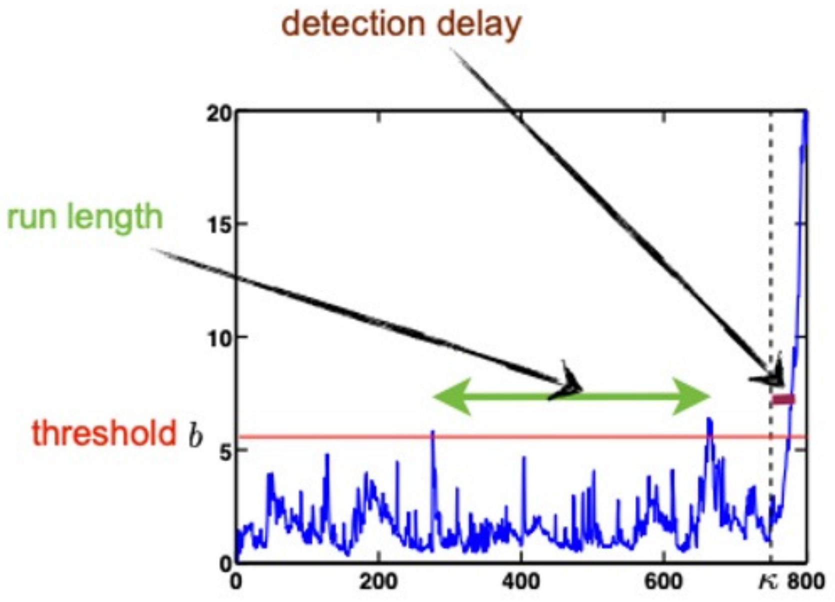
- CUSUM are **exact-optimal**

$$\mathrm{EDD}(T_C) = \min_{T:ARL(T)\geq\gamma} \mathrm{EDD}(T) = \frac{\log\gamma}{\mathrm{I}_0}(1+o(1))$$

$\mathrm{I}_0 = KL(f_1||f_0)$ is the KL divergence.

- GLR are **asymptotically-optimal**

$$\frac{\mathrm{EDD}(T_G)}{\min_{T:ARL(T)\geq\gamma} \mathrm{EDD}(T)} \to 1, \text{ as } \gamma\to\infty, (\mathrm{ARL}(T_G)\geq\gamma)$$



- **Average run length (ARL)**

$$\mathbb{E}_\infty[T]$$

- ***Worst-case* expected detection delay (EDD)**

$$\sup_{\tau\geq 1} ess\sup \mathbb{E}_\tau[(T-\tau)^+|x_1,\ldots,x_{\tau-1}]$$

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, *14*(4), 1379-1387.
Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information theory*, *44*(7), 2917-2929.

This talk:

A new online change detection algorithm that

- **Can be used when post-change density is unknown (while CUSUM cannot)**
- **Computationally more efficient than GLR**
- **Enjoys first-order asymptotic optimality (same as GLR)**

Agenda

- Background on Sequential Change-point Detection
- A Newly Proposed Window-limited CUSUM Method
- Examples and Applications

Our Method: Window-Limited CUSUM

- Recall CUSUM statistic has recursive formulation:

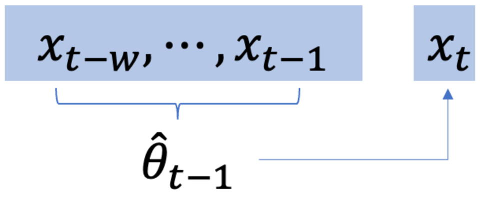
$$S_t = S_{t-1}^{+} + \log \frac{f_1(x_t)}{f_0(x_t)}, \qquad S_0 = 0.$$

- Window-limited CUSUM: Still enjoys recursive formulation

$$\mathcal{S}_t = \mathcal{S}_{t-1}^{+} + \log \frac{f_1(x_t; \widehat{\theta}_{t-1})}{f_0(x_t)}, \qquad \mathcal{S}_w = 0, t = w+1, w+2, \ldots$$

- Stopping time: $\mathcal{J} = \inf\{t \geq w : \mathcal{S}_t \geq \nu\}$
- Use *any* **consistent** estimate $\hat{\theta}_t$, e.g., the MLE

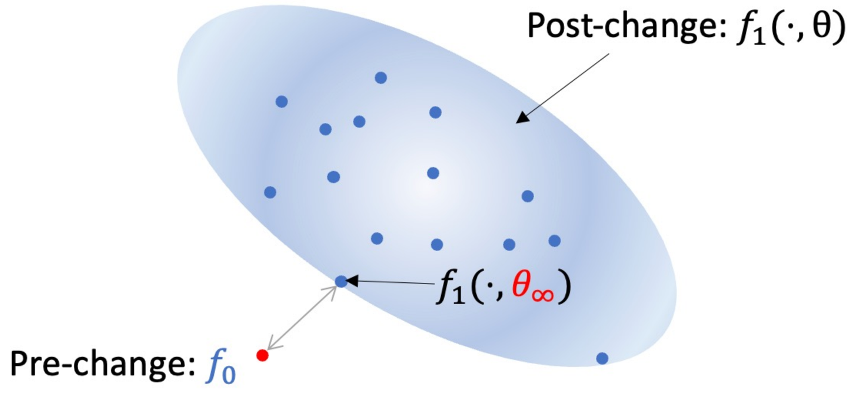
$$\hat{\theta}_t = \arg\max_{\theta \in \Theta} \sum_{i=0}^{w-1} \log f_1(x_{t-i}; \theta)$$



Xie, L., Moustakides, G. V., & Xie, Y. (2023). Window-Limited CUSUM for Sequential Change Detection. IEEE *Transactions on Information Theory*.

Key Assumptions

- ✓ The pre-change distribution f_0 does not belong to the post-change distributions $f_1(\cdot, \theta)$ with parameter set Θ.



- ✓ Under the post-change regime, we assume the estimate $\hat{\theta}_t$ is (asymptotically) unbiased. And $Cov(\hat{\theta}_t) = \frac{1}{w}\Sigma_0(1 + o(1))$.

- ✓ Under the pre-change regime, we assume the mean of the estimator $\hat{\theta}_t$ is (asymptotically) θ_∞. And $Cov(\hat{\theta}_t) = \frac{1}{w}\Sigma_\infty(1 + o(1))$.

$$\theta_\infty = \arg\max_\theta \mathsf{E}_\infty[\log \mathsf{f}_0(\xi_1, \theta)]$$

Theoretical Guarantee - ARL

Lemma (ARL of the proposed method) $\mathrm{ARL}(\mathcal{J}) \geq e^{\nu} = \gamma$ $\Rightarrow$ $\nu = \log \gamma$

Proof sketch: Define a new statistics

$$\mathcal{L}_t = (\mathcal{L}_{t-1} + 1)\frac{\mathsf{f}_0(\xi_t, \hat{\theta}_{t-1})}{\mathsf{f}_\infty(\xi_t)}, \ \ \mathcal{L}_w = 0.$$

$$\mathcal{L}_t \geq e^{\mathcal{S}_t} \quad \longleftrightarrow$$

Window-limited CUSUM

$$\mathcal{S}_t = \mathcal{S}_{t-1}^{+} + \log \frac{\mathsf{f}_0(\xi_t, \hat{\theta}_{t-1})}{\mathsf{f}_\infty(\xi_t)}, \mathcal{S}_w = 0.$$

$$\Downarrow$$

$$\mathsf{E}_\infty[\mathcal{L}_t - t|\mathscr{F}_{t-1}] = \mathsf{E}_\infty\left[(\mathcal{L}_{t-1}+1)\frac{\mathsf{f}_0(\xi_t, \hat{\theta}_{t-1})}{\mathsf{f}_\infty(\xi_t)} - t\middle|\mathscr{F}_{t-1}\right] = \mathcal{L}_{t-1} - (t-1).$$

Martingale

$$\Downarrow$$

$$\mathsf{E}_\infty[\mathcal{L}_T - T] = \mathsf{E}_\infty[\mathcal{L}_w - w] = -w \Rightarrow \mathsf{E}_\infty[T] - w = \mathsf{E}_\infty[\mathcal{L}_T].$$

Optional stopping

$$e^{\nu} \leq \mathsf{E}_\infty[e^{\mathcal{S}_\mathcal{T}}] \leq \mathsf{E}_\infty[\mathcal{L}_\mathcal{T}] = \mathsf{E}_\infty[\mathcal{T}] - w \leq \mathsf{E}_\infty[\mathcal{T}]$$

Theoretical Guarantee - EDD

Theorem 1 (EDD of the proposed method)

Assume $\hat{I}_0$>0. We can upper bound the worst-case EDD of the proposed window-limited CUSUM:

$$\mathrm{EDD}(\mathcal{J}) \leq \frac{\overset{\boldsymbol{\nu}}{\log\gamma} + \frac{\hat{J}_0}{\hat{I}_0} + \left(\frac{\hat{J}_0}{\hat{I}_0}\log\gamma\right)^{\frac{1}{2}} + wI_0 + \left(\frac{\hat{J}_0}{\hat{I}_0}I_0 w\right)^{\frac{1}{2}}}{\hat{I}_0}$$

where $\hat{I}_0 = \mathbb{E}_0\left[\log\frac{f_1(x_t;\hat{\theta}_{t-1})}{f_0(x_t)}\right], \hat{J}_0 = \mathbb{E}_0\left[\left(\log\frac{f_1(x_t;\hat{\theta}_{t-1})}{f_0(x_t)}\right)^2\right]$

- As γ is usually large, the upper bound is dominated by $\frac{\log\gamma}{\hat{I}_0}$
- Recall the EDD of the exact-CUSUM is $\frac{\log\gamma}{I_0}(1+o(1))$

Theoretical Guarantee - EDD

$$\mathrm{EDD}(\mathcal{J}) \leq \frac{\log\gamma + \frac{\hat{J}_0}{\hat{I}_0} + \left(\frac{\hat{J}_0}{\hat{I}_0}\log\gamma\right)^{\frac{1}{2}} + wI_0 + \left(\frac{\hat{J}_0}{\hat{I}_0}I_0 w\right)^{\frac{1}{2}}}{\hat{I}_0}$$

Proof sketch:

1. We can show that $\sup_{\tau\geq 1} ess\sup \mathbb{E}_\tau[(\mathcal{J}-\tau)^+ | x_1, \ldots, x_{\tau-1}]$ = $\mathbb{E}_0[\mathcal{J}]$ **(The worst-case delay is attained when change-point is 0)**

2. Relate to another stopping rule: $\mathcal{U}_t = \mathcal{U}_{t-1} + \log\frac{f_1(x_t;\ \hat{\theta}_{t-1})}{f_0(x_t)}$, $\mathcal{U}_w = 0.$ $\quad \mathcal{J}' = \inf\{t \geq w: \mathcal{U}_t \geq \nu\}.$

Original: $\mathcal{S}_t = \mathcal{S}_{t-1}^+ + \log\frac{f_1(x_t;\ \hat{\theta}_{t-1})}{f_0(x_t)}$

3. Bound the detection delay of the alternative stopping time $\mathcal{J}'$, using **Wald's identity for w-dependent sequences.**

$$\mathbb{E}_0[\mathcal{J}'] \leq \frac{\mathbb{E}_0[\mathcal{U}_{\mathcal{J}'}] + wI_0}{\hat{I}_0} = \frac{\mathbb{E}_0[\mathcal{U}_{\mathcal{J}'} - \nu] + \nu + wI_0}{\hat{I}_0}$$

4. Bound the **overshoot**

Theoretical Guarantee - EDD

$$\mathrm{EDD}(\mathcal{J}) \leq \frac{\log\gamma + \frac{\hat{J}_0}{\hat{I}_0} + \left(\frac{\hat{J}_0}{\hat{I}_0}\log\gamma\right)^{\frac{1}{2}} + wI_0 + \left(\frac{\hat{J}_0}{\hat{I}_0} I_0 w\right)^{\frac{1}{2}}}{\hat{I}_0}$$

Proof sketch:

Main Technique 1: Wald's identity for w-dependent sequences $x_1, x_2, \dots, x_n, x_{n+1}, \dots, x_{n+w}, x_{n+w+1}, \dots,$

independent

$$\Rightarrow \quad \hat{\mathsf{I}}_0 \mathsf{E}_0^\theta[\mathcal{T}'] - w\mathsf{I}_0 \leq \mathsf{E}_0^\theta[\mathcal{U}_{\mathcal{T}'}] \quad \Rightarrow \quad \mathsf{E}_0^\theta[\mathcal{T}'] \leq \frac{\mathsf{E}_0^\theta[\mathcal{U}_{\mathcal{T}'}] + w\mathsf{I}_0}{\hat{\mathsf{I}}_0} = \frac{\mathsf{E}_0^\theta[\mathcal{U}_{\mathcal{T}'} - \nu] + \nu + w\mathsf{I}_0}{\hat{\mathsf{I}}_0}$$

Main Technique 2: Bounding the Overshoot $\mathcal{R}_\nu = \mathcal{U}_{\mathcal{T}'} - \nu$

$$\left(\mathsf{E}_0^\theta[\mathcal{R}_\nu]\right)^2 \leq \frac{\hat{\mathsf{J}}_0}{\hat{\mathsf{I}}_0}\left\{\mathsf{E}_0^\theta[\mathcal{R}_\nu] + \nu + w\mathsf{I}_0\right\} \quad \Rightarrow \quad \mathsf{E}_0^\theta[\mathcal{R}_\nu] \leq \frac{\hat{\mathsf{J}}_0}{\hat{\mathsf{I}}_0} + \left(\frac{\hat{\mathsf{J}}_0}{\hat{\mathsf{I}}_0}\nu\right)^{\frac{1}{2}} + \left(\frac{\hat{\mathsf{J}}_0}{\hat{\mathsf{I}}_0}\mathsf{I}_0 w\right)^{\frac{1}{2}}$$

G. V. Moustakides, "Extension of Wald's first lemma to Markov processes," J. Appl. Probab., vol. 36, no. 1, pp. 48–59, 1999.
G. Lorden, "On excess over the boundary," Ann. Math. Statist., vol. 41, no. 2, pp. 520–527, 1970.

Theoretical Guarantee - Optimality

Theorem 2 (Asymptotic optimality)

If the window size w is chosen as $w = \underbrace{O\big((\log\gamma)^{\alpha}\big)}^{o(\log\gamma)}$ with $0 < \alpha < 1$ then

$$\mathbf{EDD}(\mathcal{J})/\frac{\log\gamma}{I_0} \le 1 + O\left(\frac{1}{(\log\gamma)^{\beta}}\right), \quad \beta = \min\left\{\alpha, 1-\alpha, \frac{1}{2}\right\}.$$

$(\to 1, as\ \gamma \to \infty)$

The maximal convergence to unity is achieved when $\beta = \frac{1}{2}$, which occurs for $\alpha = \frac{1}{2}$: $w = O\big(\sqrt{\log\gamma}\big)$.

$$1 = \lim_{\gamma\to\infty} \mathbb{E}_0[T_C]/\frac{\log\gamma}{I_0} \le \lim_{\gamma\to\infty} \mathbb{E}_0[\mathcal{J}]/\frac{\log\gamma}{I_0} \le \lim_{\gamma\to\infty} \left\{1 + \boxed{O\left(\frac{1}{(\log\gamma)^{\beta}}\right)}\right\} = 1$$

CUSUM

WLCUSUM (proposed)

Optimal window-size:

$$w = \frac{\sqrt{\mathrm{trace}\{\Sigma_0 \mathrm{F}_0\}}}{\mathrm{I}_0\sqrt{2}} \sqrt{\log\gamma}\left\{1 + \Theta\left(\frac{1}{(\log\gamma)^{\frac{1}{4}}}\right)\right\}$$

Recall GLR requires window size of the order $\boldsymbol{O(\log\gamma)}$

Comparison of Complexity and Optimality

Exact-CUSUM	Window-limited GLR	WL-CUSUM
$S_t = S_{t-1}^+ + \log\frac{f_1(x_t)}{f_0(x_t)}$	$S_t^{wGLR} = \max_{t-w\le k<t} \sup_{\theta\in\Theta} \sum_{i=k+1}^{t} \log\frac{f_1(x_i;\theta)}{f_0(x_i)}$	$\mathcal{S}_t = \mathcal{S}_{t-1}^+ + \log\frac{f_1(x_t;\,\hat{\theta}_{t-1})}{f_0(x_t)}$
$\mathbb{E}_0[\mathrm{T_c}] = \frac{\log\gamma}{\mathrm{I}_0}\left(1+O\left(\frac{1}{\log\gamma}\right)\right)$	$\mathbb{E}_0[\mathrm{T_G}] \le \frac{\log\gamma}{\mathrm{I}_0}(1+o(1))$	$\mathbb{E}_0[\mathcal{J}] \le \frac{\log\gamma}{I_0}\left(1+O\left(\frac{1}{(\log\gamma)^\beta}\right)\right)$
No window size, most **efficient** But require f_0, f_1 **known**	Window size $\liminf w/\log\gamma > \mathrm{I}_0^{-1}$ and $\log w = o(\log\gamma)$	Window size only $w = O\left((\log\gamma)^\alpha\right) \quad 0<\alpha<1$

Complexity $O\left((\log\gamma)\right)^2$ per time update Complexity $O\left(\sqrt{\log\gamma}\right)$ per time update

WL-CUSUM is more efficient, and still asymptotic optimal

A Variant – Parallel WLCUSUM

Some **limitations** of the proposed WLCUSUM procedure:

- For any fixed w, the WLCUSUM procedure will start after observing w samples, meaning a delay at least w

- The optimal window size is unknown when the post-change signal strength I_0 is unknown

$$w = \frac{\sqrt{\text{trace}\{\Sigma_0 \mathsf{F}_0\}}}{\mathsf{I}_0\sqrt{2}}\sqrt{\log\gamma}\left\{1+\Theta\left(\frac{1}{(\log\gamma)^{\frac{1}{4}}}\right)\right\}$$

Solution: parallel WLCUSUMS with different window sizes, ranging from 1 to max W.

Denote $\mathcal{J}(w)$ as the stopping time using window size w.

Parallel approach's stopping time: $\mathcal{J}_P = \min_{1\le w\le W} \mathcal{J}(w)$

- we **do not wait** for W samples to start the detection procedure
- we **do not need to specify an exact window size** beforehand
- With a suitable choice of W, such parallel procedure is still **asymptotic optimal**

Examples: subspace change detection

- **Emerging subspace** problem:

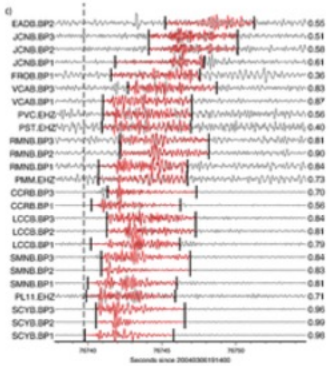
$$
\begin{aligned}
x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_p), & t &= 1, 2, \ldots, {\color{red}\tau}, \\
x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_p + {\color{red}\theta u u^\intercal}), & t &= \tau + 1, \ldots
\end{aligned}
$$

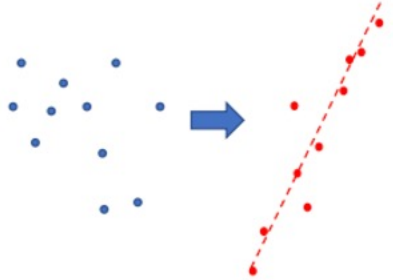
- Unknown τ, θ, u.
- Proposed **Subspace-CUSUM** Statistics:

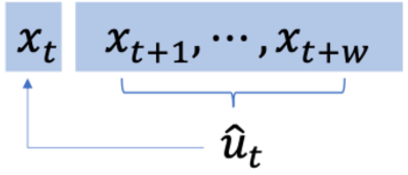
$$
\begin{aligned}
\mathcal{S}_0 &= 0, \\
\mathcal{S}_t &= (\mathcal{S}_{t-1})^+ + \log \frac{{\color{red}f_1}(x_t)}{{\color{blue}f_0}(x_t)} \\
&\approx (\mathcal{S}_{t-1})^+ + ({\color{red}\hat{u}_t}^\intercal x_t)^2 - \underbrace{\sigma^2(1 + \sigma^2/\theta)\log(1 + \theta/\sigma^2)}_{\color{red}d}
\end{aligned}
$$

$$
\mathcal{T} = \inf\{t : \mathcal{S}_t > b\}.
$$

dense geophysical sensor array





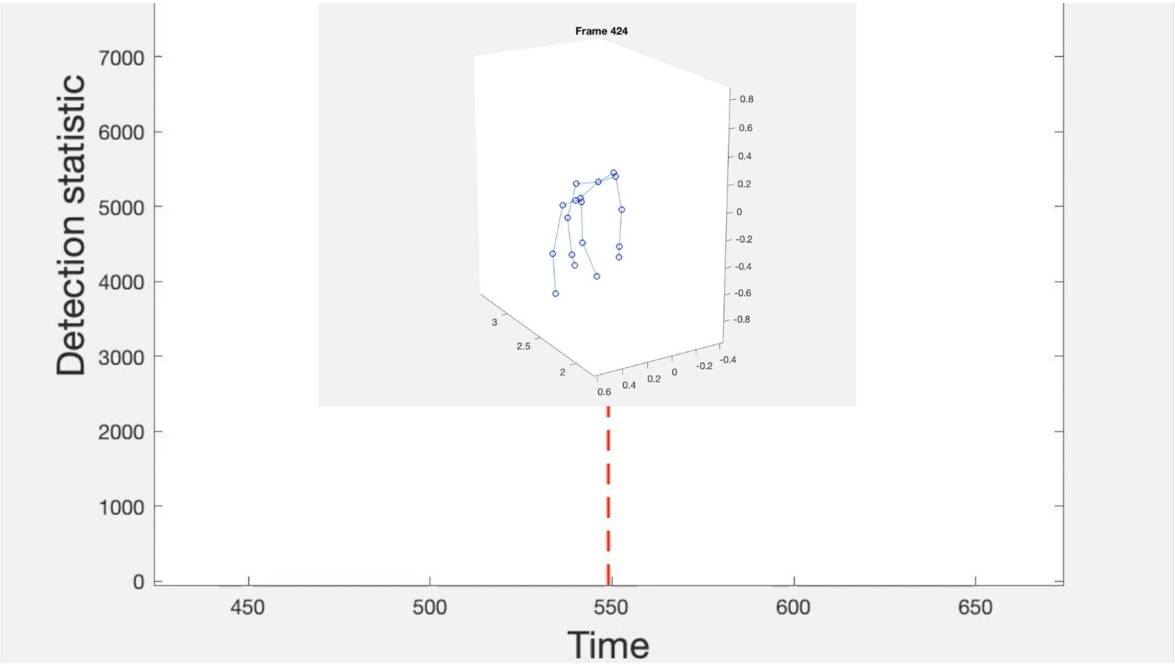


Many ***online subspace tracking*** algorithms: GROUSE, PETRELS, PAST, ……

Xie, L., Xie, Y., & Moustakides, G. V. (2020). Sequential subspace change point detection. *Sequential Analysis*, *39*(3), 307-335.

Xie, L., Moustakides, G. V., & Xie, Y. (2018). First-order optimal sequential subspace change-point detection. *GlobalSIP* (pp. 111-115). IEEE.

Examples: subspace change detection (rank-k)

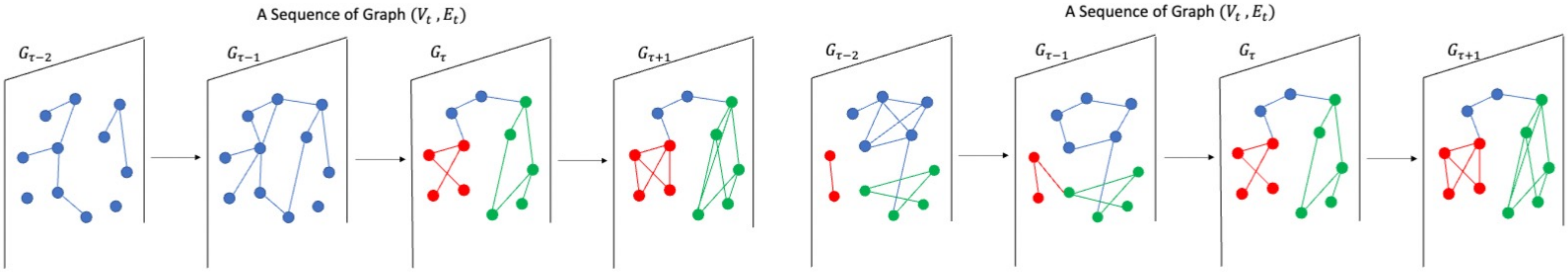


- Real-time location of 18 body joints
- Each joint is a point in the Cartesian coordinate, changing over time
- Dynamic (gesture-dependent) correlation among different joints

Microsoft Research Cambridge-12 Kinect gesture dataset. https://www.microsoft.com/en-us/download/details.aspx?id=52283

Examples: community detection

n nodes
$m(\ll n)$ communities



(a) Emergence problem illustration (b) Switching membership problem illustration

$$
\begin{aligned}
H_0 : \quad & v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{\sigma^2} I), \quad t = 1, 2, \ldots \\
H_1 : \quad & v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{\sigma^2} I), \quad t = 1, 2, \ldots, \tau, \\
& v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, (AA^\top + \sigma^2 I)^{-1}), \quad t = \tau + 1, \tau + 2, \ldots
\end{aligned}
$$

Inverse covariance matrix
Gaussian graphical mode

For example, $n = 3, m = 2$

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

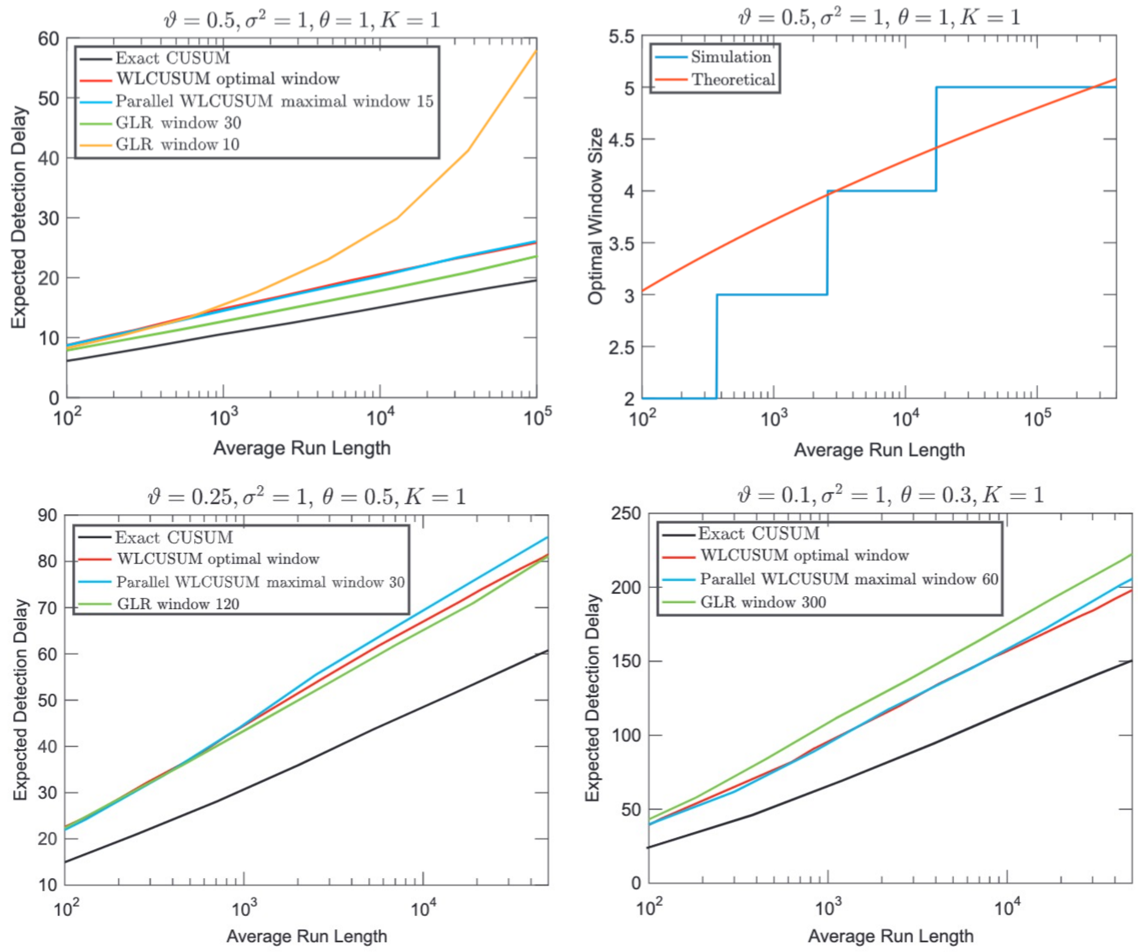
Node-1 belongs to 1st community
Node-2 belongs to 1st community
Node-3 belongs to 2nd community

$$AA^T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$ *adjacency matrix*

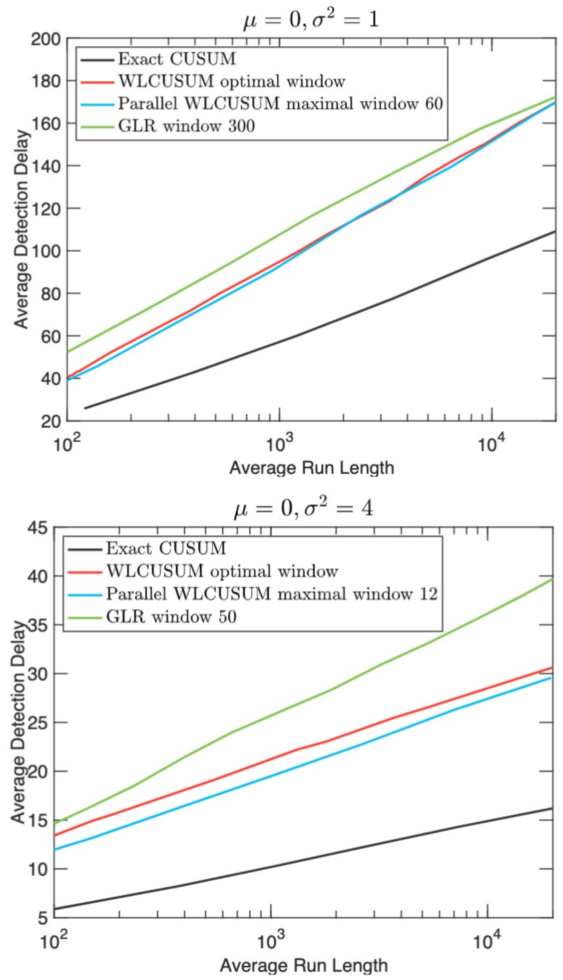
Estimate unknown A using eigen-decomposition, or subspace tracking, …

Zhang, M., Xie, L., Xie, Y. (2023). Spectral CUSUM for Online Network Structure Change Detection. *IEEE Transactions on Information Theory.*

Numerical Demonstrations:



Univariate Normal mean-shift



Laplace to normal

WL-CUSUM performs **better** than GLR when the signal strength is weak, with much **less** operations

Summary and Discussion

- A new CUSUM type approach for online change detection
  - ➢Joint detection & estimation, for unknown post-change densities
  - ➢Computationally efficient
  - ➢First-order asymptotic optimality
- Opportunities and future directions
  - online estimate of unknown post-change parameters (gradient descent, etc.)
  - non-parametric and data-driven problems
  - Robust detection (robust to distributional uncertainties, data contamination, etc.)

Thanks!

Any questions?

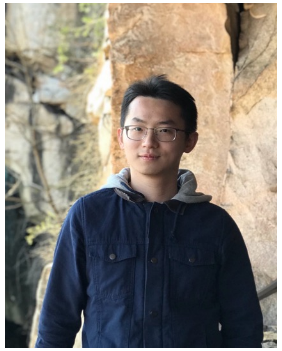
xieliyan@cuhk.edu.cn

xlyustc@gmail.com

https://liyanxie.github.io/

George V. Moustakides
University of Patras

Yao Xie
Georgia Tech



Minghe Zhang
Georgia Tech

Xie, L., Moustakides, G. V., & Xie, Y. (2023). Window-Limited CUSUM for Sequential Change Detection. *IEEE Transactions on Information Theory.*